Rethinking Reverse Distillation for Multi-Modal Anomaly Detection

Zhihao Gu[1*], **Jiangning Zhang**[2], **Liang Liu**[2], **Xu Chen**[2], **Jinlong Peng**[2], **Zhenye Gan**[2], **Guannan Jiang**[3], **Annan Shu**[3], **Yabiao Wang**[2], **Lizhuang Ma**[1†]

[1]School of Electronic and Electrical Engineering, Shanghai Jiao Tong University
[2]YouTu Lab, Tencent
[3]Contemporary Amperex Technology Co. Limited (CATL)
ellery-holmes@sjtu.edu.cn, {vtzhang, cxxuchen, jeromepeng, wingzygan}@tencent.com, ma-lz@cs.sjtu.edu.cn

Abstract

In recent years, there has been significant progress in employing color images for anomaly detection in industrial scenarios, but it is insufficient for identifying anomalies that are invisible in RGB images alone. As a supplement, introducing extra modalities such as depth and surface normal maps can be helpful to detect these anomalies. To this end, we present a novel ***M**ulti-**M**odal **R**everse **D**istillation* (***MMRD***) paradigm that consists of *a frozen multi-modal teacher encoder* to generate distillation targets and *a learnable student decoder* targeting to restore multi-modal representations from the teacher. Specifically, the teacher extracts complementary visual features from different modalities via a *siamese architecture* and then *parameter-freely* fuses these information from multiple levels as the targets of distillation. For the student, it learns modality-related priors from the teacher representations of normal training data and performs interaction between them to form multi-modal representations for target reconstruction. Extensive experiments show that our *MMRD* outperforms recent state-of-the-art methods on both anomaly detection and localization on MVTec-3D AD and Eyecandies benchmarks. Codes will be available upon acceptance.

Introduction

Anomaly detection (AD) has received continuous attention for several decades due to its wide range of applications such as defect detection, autonomous driving, video surveillance, and medical diagnosis. It is usually formulated as an unsupervised problem for the scarcity of anomalous data.

In recent years, vast efforts are dedicated to developing unsupervised anomaly detectors in images and tremendous progress has been made (Rudolph, Wandt, and Rosenhahn 2021; Roth et al. 2022; Li et al. 2021; Zavrtanik, Kristan, and Skočaj 2021; Hou et al. 2021; Deng and Li 2022), where embedding-based methods, synthesis and reconstruction are the dominant trends for this task. Embedding-based methods (Rudolph, Wandt, and Rosenhahn 2021; Roth et al. 2022) characterize the corresponding distribution of the extracted features, and the anomalies are detected by measuring the distance between features of test images and the estimated distribution. The synthesis-based methods (Li et al. 2021; Zavrtanik, Kristan, and Skočaj 2021) estimate the decision boundary between anomaly-free samples and the synthetically anomalous data for detection. Contrarily, methods by reconstruction (Hou et al. 2021; Deng and Li 2022) either recover the input (Hou et al. 2021) or restore middle-level features (Deng and Li 2022), as shown in Fig. 1-**Left**, where the pixel-wise similarity indicates the anomalies. However, extensive investigations in Invest3D (Horwitz and Hoshen 2022) show that some anomalies are hard to be detected on RGB images. Therefore, a few 3D-based methods are motivated to be developed, which directly deal with the 3D data for anomaly detection. For instance, Invest3D extracts orientation-invariant 3D features via FPFH (Rusu, Blodow, and Beetz 2009) operator and adopts PatchCore (Roth et al. 2022) for detection. And 3D-ST (Bergmann and Sattlegger 2023) extends the 2D teacher-student network to anomaly-free point clouds. Nevertheless, they usually produce inferior results than their RGB-based counterparts due to the complexity of 3D data. To improve the effectiveness, recent arts (Rudolph et al. 2023; Bonfiglioli et al. 2022; Wang et al. 2023) tend to utilize multiple modalities for AD, the necessity of which is illustrated in Fig. 2. The hole on the "cookies" and the protrusion on the "lollipop" is imperceptible on RGB images, but can be detected using depth and surface normals as the auxiliary modality. Besides, they also pro-

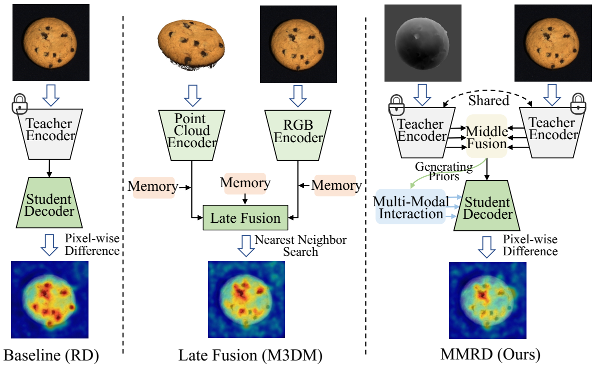


Figure 1: Illustration of different multi-modal anomaly detectors and corresponding anomaly maps (last row). Left: Reverse distillation. Middle: Two-stream structure with late fusion. Right: Our proposed paradigm.

[*]Work done when Zhihao Gu is an intern at CATL.

[†]Corresponding author.

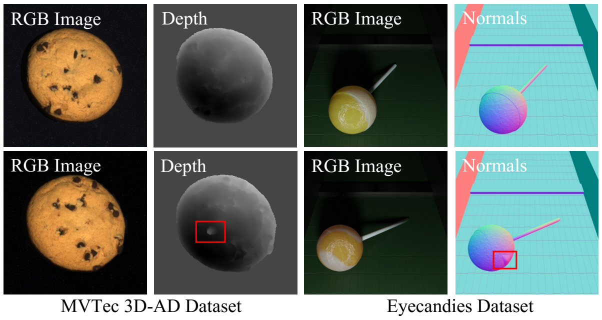


Figure 2: First row: normal samples. Second row: defective samples. Depth and normals provide supplementary visual information to RGB images for revealing anomalies and reducing misidentification of anomaly-free areas.

vide supplementary visual information reduce misidentification of anomaly-free areas. Among those methods, autoencoder (Bonfiglioli et al. 2022) is used to reconstrcut the concatenation of RGB and depth images. And M3DM (Wang et al. 2023) extends the 2D PatchCore to deal with different modalities by complicated networks and lately fuse them for multi-modal AD. However, the knowledge distillation (KD), as one of the mainstream approaches in 2D AD, has not been explored. A natural queston is: *how can we develop an efficient KD paradigm from a multi-modal perspective?*

This paper answers it in the context of Reverse Distillation (RD) (Deng and Li 2022) and presents a novel ***M**ulti-**M**odal **R**everse **D**istillation* (***MMRD***) paradigm for multi-modal anomaly detection. The main idea is to integrate information in the auxiliary modality to the frozen teacher encoder and learnable student decoder at multiple feature levels (Fig. 1-**Right**). The resulting multi-modal teacher encodes supplementary information from the auxiliary modality via a *siamese structure*, and *parameter-freely* fuses the RGB features with them as the multi-modal targets of distillation. Instead the multi-modal student *learns modality-related priors* from the normal data during training and *interactively produces multi-modal representations* to restore those targets. Consequently, the proposed *MMRD* achieves state-of-the-art results on two multi-modal AD benchmarks. What's more, it is not only flexible, handling images, depth, and surface normals but also generalizable to another distillation paradigm, *i.e.,* the forward distillation (Bergmann et al. 2020). To sum up, our main contributions are fourfold:

- We develop a novel reverse distillation paradigm, named *MMRD*, for multi-modal anomaly detection.
- We devise a frozen multi-modal teacher encoder to generate multi-modal distillation targets through a siamese structure and a parameter-free modulation module.
- We design a learnable multi-modal student decoder to restore representations of the multi-modal teacher via generating multi-modal priors.
- The proposed *MMRD* achieves state-of-the-art results on two multi-modal anomaly detection benchmarks.

Related Work

Unsupervised Anomaly Detection. Most existing works detect anomalies on RGB images and can be classified into three categories (Xie et al. 2023a): synthesis-based (Li et al. 2021; Zavrtanik, Kristan, and Skočaj 2021), embedding-based (Rudolph, Wandt, and Rosenhahn 2021; Roth et al. 2022; Gu et al. 2023; Xie et al. 2023b) and reconstruction-based (Hou et al. 2021; Deng and Li 2022; Liang et al. 2023) methods. Contrarily, limited methods perform unsupervised 3D anomaly detection (Liu et al. 2023; Chen et al. 2023a). Grid-VAE (Bengs et al. 2021) adopts the variational Auto-Encoder (AE) to reconstruct 3D voxel grids and produces anomaly scores by comparing each voxel element of the input to its reconstruction. 3D-ST (Bergmann and Sattlegger 2023) adapts the 2D student-teacher framework to detect geometric anomalies in high-resolution 3D point clouds. However, they do not perform well on the challenging MVTec 3D-AD (Bergmann et al. 2022) benchmark due to the complexity of 3D data, and new methods are needed. Recent efforts tend to combine different modalities for better anomaly detection. AST (Rudolph et al. 2023) proposes an asymmetric student-teacher network to deal with the concatenation of image features and depth maps. Eyecandy (Bonfiglioli et al. 2022) directly concatenates different modalities along the channel dimension as the input and reconstructs it via an AE. M3DM (Wang et al. 2023) uses a two-stream structure to extract features from different modalities and lately fuses them for AD. Two aspects differ our method from the above ones: 1) we develop a novel multi-modal reverse distillation paradigm and 2) integrate features of different modalities at multiple feature levels.

Knowledge Distillation. Knowledge distillation (Hinton, Vinyals, and Dean 2015; Gou et al. 2021, 2022) is originally used to transfer knowledge from a heavy teacher model to a lightweight student network and has achieved prominent progress in many fields. In AD, the student tends to unsuccessfully reconstruct the features of the teacher for anomalous samples. This insight is used to localize anomalies. US (Bergmann et al. 2020) first introduces KD for the task. Later, the forward distillation (Salehi et al. 2021; Wang et al. 2021) forms the Student-Teacher (S-T) feature pyramid and performs multi-scale feature distillation. Differences between multi-features are exploited for localization. However, the RD (Deng and Li 2022) argues that similar structures between the S-T harm the feature diversity and thus the student is built on top of the teacher. All these methods have difficulty in handling anomalies invisible in RGB images. Instead, for the first time, we explore KD to deal with these anomalies from a multi-modal perspective.

Multi-Modal Fusion. Different modalities contain supplementary information and fusing them is beneficial to better understand visual scenes compared to methods with one modality as input (Liu et al. 2022; Zhang et al. 2023b; Chen, Han, and Zhang 2023; Chen et al. 2023b; Zhang et al. 2023a). CEN (Wang et al. 2020a) dynamically exchanges the channels between sub-networks for fusion based on the scaling factor of the batch normalization. Asym-Fusion (Wang et al. 2020b) performs asymmetric shuffle

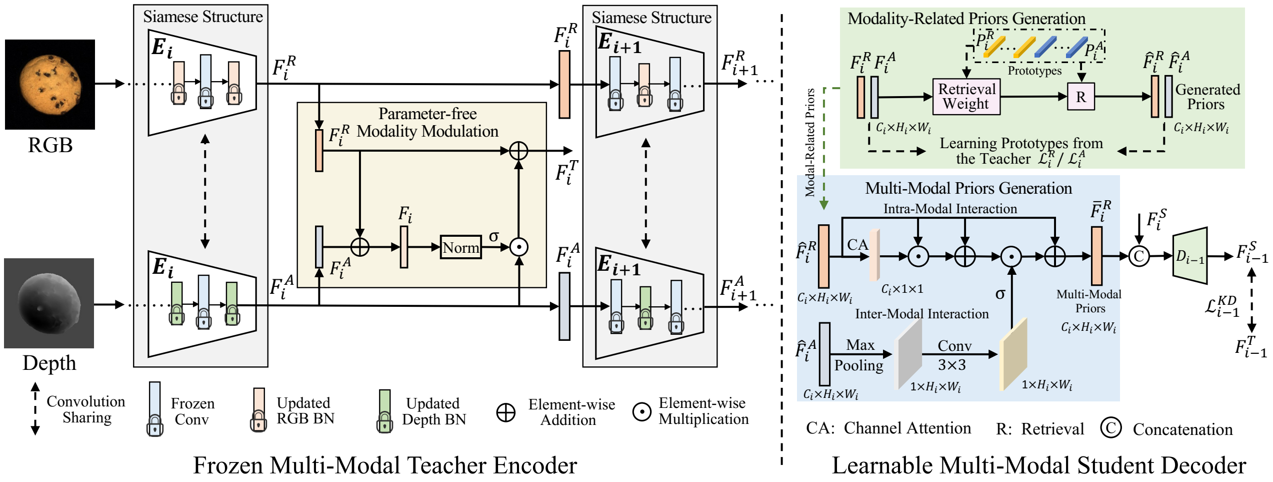


Figure 3: Overview of the proposed multi-modal reverse distillation (MMRD). It comprises a frozen multi-modal teacher encoder and a learnable multi-modal student decoder, and each of them contains two important components. At i^{th} stage, the teacher adopts a *siamese encoder* E_i with frozen convolutions and individual BNs to extract supplementary visual information, *i.e.,* F_i^R and F_i^A, from RGB image and the auxiliary modality. A *parameter-free modality modulation module* then fuses them and produces the distillation target F_i^T. Instead, the student *generates modality-related priors*, *i.e.,* $\hat{F}_i^R$ and $\hat{F}_i^A$, by learning prototypes, *i.e.,* P_i^R and P_I^A, from the teacher representations of normal data, *i.e.,* F_i^R and F_i^A, and then performs interaction between $\hat{F}_i^R$ and $\hat{F}_i^A$ to *generate multi-modal representation* $\bar{F}_i^R$. Finally, $\bar{F}_i^R$ is concatenated with the student representation F_i^S to restore target F_i^T. In inference, pixel-wise similarity between $\{F_i^T, F_i^S\}_{i=1}^K$ is computed for anomaly detection.

and shift operations to exchange information between multi-modal features. MGAF (Kim, Jones, and Hager 2021) fuses motion features with that from detection via the cross-attention (Wang et al. 2018). In KD for multi-modal AD, we not only perform parameter-free modality modulation to form distillation targets in the teacher but also generate multi-modal representations to help the student better restore these targets.

Proposed Method

This section revisits knowledge distillation for anomaly detection as preliminaries. Then, details of the proposed frozen multi-modal teacher encoder and learnable multi-modal student decoder are presented one by one. The overall paradigm is shown in Fig. 3 and the algorithm table summarizing the proposed method is included in the supplementary material.

Preliminaries: Knowledge Distillation for AD

In AD, the Knowledge Distillation (KD) detects anomalies based on RGB images and contains a pre-trained teacher network and a learnable student network. It owns two types: ***1)*** the Forward Distillation (FD) (Bergmann et al. 2020; Wang et al. 2021); ***2)*** Reverse Distillation (RD) (Deng and Li 2022). Formally, given an RGB image $I^R \in \mathbb{R}^{C \times H \times W}$ (C, H, and W is the channel, height, and width), the frozen teacher extracts feature $\{F_i^R\}_{i=1}^K \in \mathbb{R}^{C_i \times H_i \times W_i}$ (distillation targets) from its K stages and the student is trained to restore them, resulting in $\{F_i^S\}_{i=1}^K \in \mathbb{R}^{C_i \times H_i \times W_i}$. Differently, the student in FD encodes I^R but the student in RD decodes the one-class embedding of the teacher. Finally, a KD loss is used to supervise the reconstruction process:

$$\mathcal{L}_i^{\mathrm{KD}} = 1 - \frac{\mathrm{flat}(F_i^R)}{\|\mathrm{flat}(F_i^R)\|_2} \cdot \frac{\mathrm{flat}(F_i^S)^\mathsf{T}}{\|\mathrm{flat}(F_i^S)\|_2}, \qquad (1)$$

where $\mathrm{flat}(\cdot)$ is the flatten function. In inference, pixel-wise cosine similarity between $\{F_i^R, F_i^S\}_{i=1}^K$ is computed to detect and localize anomalies, as shown in Fig. 1-**Left**.

However, it is difficult for KD to detect anomalies invisible in RGB images. To handle it, based on RD, we develop a novel multi-modal reverse distillation paradigm, which contains a frozen multi-modal teacher encoder and a learnable multi-modal student decoder.

Multi-Modal Teacher (MMT) Encoder

For the teacher encoder, we generate multi-modal distillation targets by integrating supplementary information from an auxiliary modality with an RGB image. As illuminated in Fig. 3-**Left**, we adopt a cross-statistics siamese teacher network to extract those information and a modality modulation module to parameter-freely produce these targets.

Cross-Statistics Siamese Teacher Network. Fig. 2 shows that auxiliary modalities provide supplementary visual information to RGB images for revealing anomalies and reducing misidentification of anomaly-free areas. To model such supplementarity, we adopt a shared encoder, known as the siamese network, to extract features from the RGB image and the corresponding auxiliary modality, denoted as $\{F_i^R, F_i^A\}_{i=1}^K$. Nevertheless, the teacher network in KD is pre-trained on RGB images, and statistics stored in Batch Normalization layers (BNs) are shifted for the auxiliary modality. To mitigate this issue, we share the frozen convolutions for both modalities but maintain individual BNs for the auxiliary modality. Relevant statistics in these BNs are updated within several epochs with parameters of affine transformation unchanged, whose impacts are explored in Tab. 2 (b) and visualized in the supplementary material. In practice, we also adopt this strategy for RGB images. As a result, the extracted features are more modality-specific.

Parameter-Free Modality Modulation. Note that since the frozen teacher in KD provides deterministic distillation targets for a given input, the modality fusion should contain no learnable parameters. Besides, as discussed before, the auxiliary modality owns supplementary visual information to RGB images and is integrated for an auxiliary purpose. Therefore, not all information in F_i^A is equally needed. To this end, we propose to estimate a fusion weight for F_i^A to decide how much information is needed to be fused and then compensate F_i^R with the selected information in a residual form. Concretely, we first exploit a normalization operation to generate the fusion weight $\alpha_i^A \in \mathbb{R}^{C_i \times H_i \times W_i}$:

$$\alpha(F_i^A) = \text{Sigmoid}(\frac{(F_i^A - \mu_i^A)^2}{(\sigma_i^A)^2 + 10^{-4}}), \tag{2}$$

where $\mu_i^A = \frac{1}{H_i W_i}\sum F_i^A$ and $(\sigma_i^A)^2 = \frac{1}{H_i W_i}\sum(F_i^A - \mu_i^A)^2$. Intuitively, the normalization operation helps reduce the disturbance from modality-specific information and better reflects the position-wise intensity. In practice, we find that α_i^A calculated from the sum of F_i^R and F_i^A, denoted as F_i, performs better than $\alpha(F_i^A)$. It may be because F_i contains more comprehensive information than individual ones and thus is a better indicator for the fusion weight. We give the visual effects in Fig. 4. Finally, the multi-modal teacher representation (distillation target) F_i^T is formulated as:

$$F_i^T = F_i^R + \alpha(F_i) \cdot F_i^A. \tag{3}$$

$\alpha(F_i) \in [0, 1]$ flexibly controls the multi-modal information. Compared to F_i^R, the multi-modal F_i^T pays more attention to objects and suppresses the effects from the background, which is investigated in the supplementary material.

Analysis. The devised siamese teacher encoder differs from AsymFusion (Wang et al. 2020b) in two aspects. First, ours extracts modality-specific features by a frozen architecture but their fully learnable structure instead encodes multi-modal features in each branch. Second, we parameter-freely fuse features of each modality to generate multi-modal distillation targets while they fuse features for further encoding.

Multi-Modal Student (MMS) Decoder

For the student, we incorporate multi-modal prior information to help restore distillation targets. To this end, we first generate priors for each modality via a modality-related priors generation module and then perform interaction on them to produce multi-modal priors via a multi-modal priors generation module, as shown in Fig. 3-**Right**.

Modality-Related Priors Generation. In KD, the student is expected to restore representations of the teacher encoder. Therefore, introducing information from the teacher to the student is helpful for the reconstruction. We then propose to learn a set of representative features (named "prototypes") from the teacher representations of normal training data and generate modality-related priors to provide finer modal information. The prototypes are learned for both modalities and integrated via feature retrieval to generate priors for each modality. Formally, given the teacher representation of an RGB image $F_i^R \in \mathbb{R}^{C_i \times H_i \times W_i}$ and N prototypes $P_i^R = \{(P_i^R)_j \in \mathbb{R}^{C_i}\}_{j=1}^N$, the position-wise retrieval weight $W_i^R \in \mathbb{R}^{N \times H_i \times W_i}$ is measured as follows:

$$(W_i^R)_{j,h,w} = \frac{\exp(d((F_i^R)_{h,w}, (P_i^R)_j))}{\sum_{j=1}^N \exp(d((F_i^R)_{h,w}, (P_i^R)_j))}, \tag{4}$$

where (w, h) denotes spatial index and $\text{d}(\cdot, \cdot)$ is the cosine similarity. Aggregating P_i^R with weights at each location of W_i^R gives the reconstruction result $\hat{F}_i^R$:

$$(\hat{F}_i^R)_{w,h} = \sum\nolimits_j (W_i^R)_{j,h,w} \cdot (P_i^R)_j. \tag{5}$$

To ensure P_i^R learns representative information, we propose to enforce the similarity between the teacher representation F_i^R and the reconstruction $\hat{F}_i^R$ in the training phase:

$$\mathcal{L}_i^R = \frac{1}{HWC}\sum\nolimits_{h,w,c} \|F_i^R - \hat{F}_i^R\|_2^2. \tag{6}$$

Note that Eq. (6) is applied to all normal training samples. Therefore, the learned P_i^R contains normal information and is representative enough. This is why we call them "prototypes". In inference, the teacher representation F_i^R is used to generate the modality-specific priors $\hat{F}_i^R$ via Eq. (5).

For the auxiliary modality, we also learn a set of N prototypes $P_i^A = \{(P_i^A)_j \in \mathbb{R}^{C_i}\}_{j=1}^N$ via a similar process, producing the loss $\mathcal{L}_i^A$ and priors $\hat{F}_i^A$.

Multi-Modal Priors Generation. Next, we aim to provide multi-modal prior information for the student to reconstruct the distillation target F_i^T. To achieve this, we perform multi-modality interaction between the modality-related $\hat{F}_i^R$ and $\hat{F}_i^A$ to obtain a refiner representation. Since the auxiliary modality provides supplementary visual cues and the student is learnable, we use $\hat{F}_i^A$ to enhance $\hat{F}_i^R$ through the intra and inter-modal interaction, as demonstrated in Fig. 3-**Right**. Specially, we first conduct the Channel Attention (CA) (Hu, Shen, and Sun 2018) on $\hat{F}_i^R$ for intra-modal enhancement. Then the Spatial Attention (SA) map of size $\mathbb{R}^{1 \times H_i \times W_i}$ is generated from $\hat{F}_i^A$ via the $\text{MaxPooling} - \text{Conv}_{3\times3} - \text{Sigmoid}$ procedure. Finally, we perform inter-modal interaction by multiplying the enhanced $\hat{F}_i^R$ with the SA map to highlight locations of interest, resulting in a finer multi-modal representation $\bar{F}_i^R$. The whole multi-modal interaction process can be formulated as follows:

$$\bar{F}_i^R = \text{SA}(\hat{F}_i^A) \cdot (\text{CA}(\hat{F}_i^R) \cdot \hat{F}_i^R + \hat{F}_i^R) + \hat{F}_i^R. \tag{7}$$

	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
3D	FPFH	82.5/**97.3**	55.1/**87.9**	95.2/98.2	79.7/**90.6**	88.3/89.2	58.2/73.5	75.8/97.7	88.9/**98.2**	92.9/95.6	65.3/**96.1**	78.2/**92.4**
	AST	88.1/95.2	57.6/74.1	**96.5**/97.3	95.7/90.4	67.9/83.0	79.7/**83.1**	**99.0/97.8**	91.5/98.1	95.6/89.1	61.1/77.8	83.3/88.6
	M3DM	**94.1**/94.3	65.1/81.8	96.5/**97.7**	**96.9**/88.2	90.5/88.1	76.0/74.3	88.0/95.8	**97.4**/97.4	92.6/95.0	76.5/92.9	**87.4**/90.6
	Ours	82.9/92.6	**68.6**/80.6	93.7/96.5	80.4/85.8	**97.2/90.4**	**86.5**/73.1	94.7/96.2	80.6/95.8	**96.7/96.6**	**84.9**/93.6	86.6/90.1
RGB	PatchCore	87.6/90.1	88.0/94.9	79.1/92.8	68.2/87.7	91.2/89.2	70.1/56.3	69.5/90.4	61.8/93.2	84.1/90.8	70.2/90.6	77.0/87.6
	AST	94.7/85.5	92.8/90.5	85.1/80.0	**82.5**/46.6	98.1/89.4	**95.1**/52.9	89.5/83.5	61.3/54.4	99.2/87.7	82.1/60.5	88.0/73.1
	M3DM	94.4/95.2	91.8/97.2	89.6/97.3	74.9/89.1	95.9/93.2	76.7/84.3	**91.9**/97.0	64.8/95.6	93.8/96.8	76.7/96.6	85.0/94.2
	Ours	**98.7/97.0**	**93.7/98.3**	**94.3/98.2**	77.0/**92.4**	**98.1/97.6**	84.7/**87.5**	91.3/**98.1**	**75.3/97.5**	**99.3/98.4**	**85.3/97.3**	**89.8/96.2**
RGB + 3D	PatchCore	91.8/97.6	74.8/96.9	96.7/97.9	88.3/**97.2**	93.2/93.3	58.2/88.8	89.6/97.5	91.2/98.1	92.1/95.0	88.6/97.1	86.5/95.9
	AST	98.3/97.1	87.3/94.4	**97.6**/98.1	97.1/93.9	93.2/91.3	88.5/91.4	**97.4**/98.1	**98.1**/98.3	**100.0**/89.0	79.7/94.0	93.7/94.6
	M3DM	99.4/97.0	90.9/97.1	97.2/97.9	**97.6**/95.0	96.0/94.1	**94.2/93.2**	97.3/97.7	89.9/97.1	97.2/97.1	85.0/97.5	94.5/96.4
	Ours	**99.9/98.6**	**94.3/99.0**	96.4/**99.1**	94.3/95.1	**99.2/99.0**	91.2/90.1	94.9/**99.0**	90.1/**99.0**	99.4/**98.7**	**90.1/98.2**	**95.0/97.6**

(a) Anomaly detection and localization performance on the MVTec 3D-AD dataset.

	Method	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marsh.	Peppermint Candy	Mean
3D	FPFH	69.3/90.4	87.0/92.1	80.6/79.5	**92.8/94.7**	86.4/87.2	59.7/63.3	90.9/91.8	91.0/91.5	85.0/87.4	89.8/90.7	83.2/86.9
	Eyecandy	60.9/87.7	85.3/91.5	82.9/76.7	84.0/95.6	82.8/**91.0**	56.0/56.9	77.0/88.2	85.6/84.3	**91.0**/92.3	85.8/87.6	79.1/85.1
	M3DM	48.2/91.1	58.9/64.5	80.5/58.1	84.5/74.8	78.0/74.8	53.8/48.8	76.6/60.8	82.7/90.4	80.0/64.6	82.2/75.0	72.5/70.2
	Ours	**84.4/96.1**	**94.4/93.6**	**91.5/91.6**	89.4/90.2	**87.5**/88.4	**73.3/67.2**	**95.4/96.3**	**93.8/92.6**	90.1/**93.8**	**95.0/96.1**	**89.5/90.6**
RGB	PatchCore	52.5/54.3	95.4/92.8	53.4/60.1	90.7/92.4	64.6/78.2	46.6/55.7	76.2/86.0	68.2/74.5	94.4/95.3	91.5/93.3	73.4/78.3
	Eyecandy	52.7/60.7	84.8/90.4	77.2/80.5	73.4/98.2	59.0/87.1	50.8/66.2	69.3/83.6	76.0/80.5	85.1/90.7	73.0/76.2	70.1/81.4
	M3DM	**64.8**/86.7	94.9/90.4	**94.1**/80.5	**100.0**/98.2	**87.8**/87.1	63.2/66.2	**93.3**/88.2	81.1/**89.5**	**99.8**/97.0	**100.0**/96.2	**87.9**/88.0
	Ours	61.8/**91.6**	**99.5/95.2**	86.2/**83.3**	97.8/**98.3**	86.1/**87.5**	**65.8/67.2**	87.0/**88.7**	**84.0**/85.6	97.1/**97.6**	99.8/**98.5**	86.5/**89.4**
RGB + 3D	PatchCore	44.8/70.9	95.0/93.3	77.9/73.7	92.8/95.2	88.8/90.2	41.6/40.7	91.2/91.9	83.1/86.6	100.0/96.9	96.3/92.9	81.1/84.0
	Eyecandy	58.7/85.2	84.6/90.3	80.7/74.1	83.3/93.5	83.3/89.9	54.3/53.6	74.4/86.7	87.0/86.4	94.6/94.5	83.5/84.3	78.4/83.9
	M3DM	62.4/90.6	95.8/92.3	**95.8**/80.3	**100.0**/98.3	88.6/85.5	**78.5/68.8**	94.9/88.0	83.6/90.6	100.0/96.6	100.0/95.5	89.7/88.2
	Ours	**85.4/97.5**	**100.0/97.0**	94.6/**94.2**	99.8/**98.5**	**90.8/91.7**	74.7/68.0	**96.6/97.0**	**98.4/94.1**	**100.0/99.0**	**100.0/99.2**	**94.0/93.6**

(b) Anomaly detection and localization performance on the Eyecandies dataset.

Table 1: Quantitative results on (a) MVTec 3D-AD and (b) Eyecandies datasets. We report Image-level AUROC (%) ↑/Pixel-level PRO (%) ↑ and highlight methods achieving the best results in bold.

Finally, $\bar{F}_i^R$ is concatenated with F_i^S as the input of the student decoder D_{i-1} to restore F_{i-1}^T, resulting in F_{i-1}^S:

$$F_{i-1}^S = D_{i-1}([F_i^S; \bar{F}_i^R]). \tag{8}$$

The F_{i-1}^S and F_{i-1}^T are used to compute the distillation loss in Eq. (1) during training and detect anomalies in inference.

Analysis. We give some theoretical explanations on scores from priors. The student is trained to produce anomaly-free features and then anomaly-free areas are inside the convex combination of "prototypes". Finally, anomalies fail to be inside the combination and features between the teacher and student have a higher reconstruction error. This insight is used for anomaly localization. Fig. 5 verifies the analysis.

Loss Function and Anomaly Detection

Loss Function. It consists of the distillation loss from K stages and the prototype learning loss of each modality:

$$\mathcal{L} = \sum\nolimits_{i=1}^{K} \mathcal{L}_i^{\mathrm{KD}} + \lambda \sum\nolimits_{i=1}^{K} (\mathcal{L}_i^R + \mathcal{L}_i^A), \tag{9}$$

where $K = 3$ and λ is the balance factor, set 0.1 by default.

Anomaly Detection. In inference, pixel-wise cosine similarity between $\{F_i^T, F_i^S\}_{i=1}^K$ is computed and then a bilinear up-sampling operation $\mathrm{Up}(\cdot)$ is conducted to generate an anomaly map S_i. The final anomaly map A is given by:

$$A = g(\sum\nolimits_i \mathrm{Up}(1 - \mathrm{d}(F_i^T, F_i^S))), \tag{10}$$

where $g(\cdot)$ denotes the Gaussian filter (Roth et al. 2022). A gives the localization results and a larger score on it indicates a higher probability of anomaly. We simply take its maximum value as the image-level anomaly score.

Experiments

Experimental Settings

Datasets. We conduct experiments on two multi-modal benchmarks, *i.e.,* the MVTec 3D-AD (Bergmann et al. 2022) and the Eyecandies (Bonfiglioli et al. 2022). The former contains 4,147 scans captured from 10 object categories and provides modality of RGB images and Point Clouds (PCs). The latter consists of 10 categories with 1,500 samples for each type and provides RGB images, depth maps, and surface normals. Pixel-level annotations are available in both datasets to evaluate the anomaly localization performance.

Component	ROC_{AD}	ROC_{AL}	PRO
Baseline	84.0	96.7	88.9
+MMT	91.5	97.2	91.5
+MMT+PG	92.6	97.6	92.0
All	**94.0**	**98.3**	**93.6**

(a) Study on key components.

Modality	ROC_{AD}	ROC_{AL}	PRO
None	93.2	97.6	92.8
3D	93.4	97.7	93.0
RGB	93.8	98.2	93.4
All	**94.0**	**98.3**	**93.6**

(b) Study on individual BNs.

Method	ROC_{AD}	ROC_{AL}	PRO
FD	70.8	85.6	78.0
MMFD	82.5	90.2	84.4
RD	84.0	96.7	88.9
MMRD	**94.0**	**98.3**	**93.6**

(c) Study on distillation paradigms.

Modality	ROC_{AD}	ROC_{AL}	PRO
Depth	74.5	90.8	86.2
Normals	89.5	96.0	90.6
RGB	86.5	94.5	89.4
+Depth	92.8	97.2	91.3
+Normal	94.0	98.3	93.6
All	**94.4**	**98.8**	**93.9**

(d) Study on modalities.

Method	ROC_{AD}	ROC_{AL}	PRO
CSA	91.2	97.9	92.4
SSA	92.5	98.1	91.6
$\alpha = 1$	93.0	98.1	92.8
$\alpha(F_i^A)$	93.4	98.2	93.1
$\alpha(F_i)$	**94.0**	**98.3**	**93.6**
SEM(F_i)	90.2	97.0	92.6

(e) Study on fusion strategies.

$\{N_1, N_2, N_3\}$	ROC_{AD}	ROC_{AL}	PRO
$\{0, 0, 0\}$	91.5	97.2	91.5
$\{10, 10, 10\}$	93.4	97.8	92.3
$\{10, 50, 10^2\}$	**94.2**	98.2	93.0
$\{50, 50, 50\}$	94.0	**98.3**	**93.6**
$\{10^2, 50, 10\}$	94.0	98.2	93.3
$\{10^2, 10^2, 10^2\}$	93.3	98.0	93.1

(f) Study on number of prototype.

Table 2: Ablation study on the Eyecandies dataset. "PG", "MMFD", "SEM", "CSA" and "SSA" refer to the modality-related prior generation, multi-modal forward distillation, SE module, channel, and spatial self-attention, respectively.

Baseline Methods. We compare ours with several SOTA multi-modal detectors, *i.e.,* AST (Rudolph et al. 2023) using depths and RGBs, M3DM (Wang et al. 2023) using PCs and RGBs, PatchCore (Roth et al. 2022) with FPFH (Rusu, Blodow, and Beetz 2009) using PCs and RGBs, and Eyecandy (Bonfiglioli et al. 2022) using normals and RGBs.

Evaluation Metrics. The Area Under the Receiver Operator Curve (AUROC) and Precision Recall (AUPR) are used to quantify anomaly detection and localization capacity. The Per-Region Overlap (PRO) is also adopted for localization.

Implemental Details. Images are resized into 256×256 and Adam is used as the optimizer with a learning rate of 0.001. The model is trained for 400 epochs of batch size 16. the number of prototypes is set 50. The teacher network is a pre-trained WideResNet50 and the student is the same as RD. We adopt the depth and normals as auxiliary modalities for MVTec 3D-AD and Eyecandies datasets, respectively.

Main Results

Results on the MVTec 3D-AD. Tab. 1 (a) shows experimental results for anomaly detection using 3D data, RGB images, or their combination on the MVTec 3D-AD dataset. Image-level AUROC and pixel-level PRO for all classes are reported. First, we find that by solely relying on RGB images for detection, our method outperforms all 3D-based counterparts (with improvements of 2.4% on AUROC_{AD} and 3.8% on PRO_{AL}) in terms of mean values. This is likely due to the complexity of 3D data and the limited efforts put into its development. However, it is also observed that geometric information in some targets, *e.g.,* foam and peach, play a more important role in detecting anomalies (86.5% versus 84.7% on foam, and 94.7% versus 91.3% on peach) since these anomalies are visually unperceived in the 2D view. Finally, integrating 3D information gives larger improvements.

Results on the Eyecandies. The proposed method is also evaluated on the Eyecandies dataset and image-level AUROC and pixel-level PRO for all classes are reported in Tab. 1 (b). We observe that the overall performance on normals is higher than that on RGB images. This is because the normals describe the geometric shape of the target object and some geometric anomalies that are hard to be perceived from images can thus become visually identifiable, as demonstrated in Fig. 2. Additionally, introducing the normals to images further improves the performance. Compared to methods such as AST and Eyecandy that fuse multiple modalities via concatenation, our strategy performs feature-level fusion, surpassing them by a clear margin.

Method	ROC_{AD}	ROC_{AL}	PR_{AD}	PR_{AL}	PRO	GPUH/FPS
AST†	93.7	97.5	97.4	33.7	94.6	10.4/**41.0**
M3DM†	93.6	99.2	97.7	**43.9**	96.2	12.6/0.10
Ours	**95.0**	**99.2**	**98.1**	42.1	**97.6**	**5.8**/10.2

Table 3: More comprehensive results on the MVTec 3D-AD dataset. AD and AL are short for anomaly detection and localization. † means re-implementation. "GPUH" and FPS refer to GPU hours and frame per second, respectively.

More comprehensive results. In Tab. 3, our method outperforms AST and M3DM in four out of five AD metrics. Moreover, it consumes less training time and the inference speed is 1⁄4x compared to AST and 100x compared to M3DM, demonstrating both the effectiveness and efficiency.

Ablation Study

Study on key components. We study the effectiveness of the multi-modal teacher (MMT) and two key components in multi-modal student (MMS), *i.e.,* modality-related Prior Generation (PG) and multi-modal interaction, in Tab. 2 (a). RD is the baseline. Since RGB images contain limited information for geometric anomalies, the baseline thus owns inferior results. Instead, introducing an auxiliary modality to the teacher brings large improvement (7.5% $\uparrow$ on AUROC_{AD} and 2.6% $\uparrow$ on PRO_{AL}). For the student, generating modality-related priors from normal samples and conducting multi-modal interaction give improvements of different degrees. Finally, combining them all performs best.

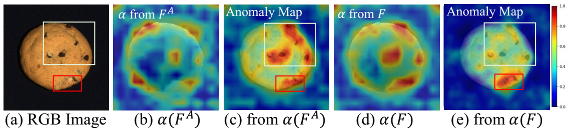
Study on individual BNs in MMT. They are used to learn modality-related statistics for adaption and their impacts are listed in Tab. 2 (b). Adopting individual BNs benefits both anomaly detection and localization while applying them to surface normals alone contributes less to final results than to RGB images, implying that the network may have difficulty further adapting Image-Net pre-trained convolutions to other modalities. Visualizations in the supplementary material show that learning RGB-related information helps the pre-trained convolutions better describe anomalies, resulting in finer multi-modal representations for the teacher.

Study on distillation paradigms. We explore the generalization of our multi-modal strategies to the Forward Distillation (FD) and Reverse Distillation (RD), as listed in Tab. 2 (c). How to apply them to FD can be found in the supplementary material. It is observed that integrating an auxiliary modality to the RGB data via our strategies gives consistent improvement to different distillation paradigms, which implies the flexibility and expandability of our method.

Study on different modalities. Tab. 2 (d) studies the effects of different modalities and how to extend our method to more modalities can be found in the supplementary material. First, compared to depth, both normals and images provide useful information for AD and thus achieve better results. Second, fusing RGB data with depth or normals all bring significant improvement whereas the normals own larger gains (6.3% *v.s.* 7.5% on AUROC$_{\text{AD}}$, 2.7% *v.s.* 3.8% on AUROC$_{\text{AL}}$ and 1.9% *v.s.* 4.2% on PRO$_{\text{AL}}$). Instead, integrating depth into images and normals produces limited improvement since depth introduces minor extra information.

Study on different fusion strategies for F_i^T. In Tab. 2 (e), we explore different ways to generate the multi-modal representation F_i^T, including the parameter-free Channel Self-Attention (CSA) and Spatial Self-Attention (SSA) (Wang et al. 2018)), and the learnable SE Module (Hu, Shen, and Sun 2018) (SEM). We observe that no learnable transformations in CSA and SSA result in inaccurate attention computation and thus lead to unsatisfactory results. Besides, they can only handle two modalities. Surprisingly, element-wise addition between F_i^R and F_i^A ($\alpha = 1$) outperforms above strategies. Contrarily, fusion with adaptive weight α produces better results, indicating that not all the information in the auxiliary modality is important. The SEM instead underperforms the vanilla addition. We guess the parameterized SEM produces unstable representations for the teacher.

Study on number of prototypes. The number of prototypes controls the amount of normal information to be learned for each modality, which is explored in Tab. 2 (f). We find that learning normal information benefits both anomaly detection and localization. And more prototypes lead to better detection while owning similar AUROC$_{\text{AL}}$. Instead, a larger N_i leads to more parameters and optimization difficulty, resulting in more performance drops. For the sake of higher localization results, we adopt $N_i = 50$ by default.



(a) RGB Image (b) $\alpha(F^A)$ (c) from $\alpha(F^A)$ (d) $\alpha(F)$ (e) from $\alpha(F)$

Figure 4: Visualization on α from different sources and corresponding detection results. Red boxes highlight anomalous areas. $\alpha(F)$ pays attention to anomalous regions and special patterns in RGB, owning more accurate localization.

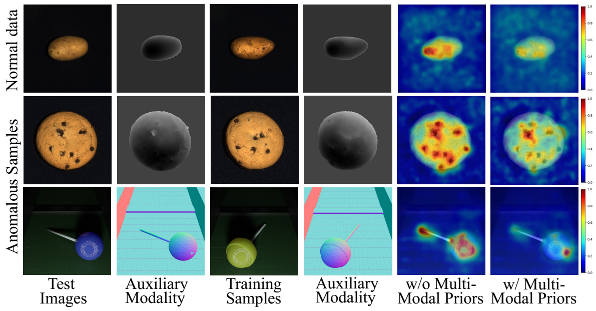


Figure 5: Visualization on multi-modal priors from training data. They help suppress sensitivity to anomaly-free patterns and give accurate localization capacity.

Visualization Analysis

Sources for generating fusion weight α. Note that α in Eq. (3) can also be obtained from F^A. To explore the difference, Fig. 4 visualizes α on the depth map and their corresponding anomaly map on the image. As shown in Fig. 4 (b) and (d), $\alpha(F)$ highlights not only anomalous regions which are visible in auxiliary modality but also some regions with special patterns in RGB (the chocolate on the "cookie"). In this sense, $\alpha(F^A)$ fails to introduce auxiliary modality information in special pattern regions and leads to wrong results in Fig. 4 (c). On the contrary, $\alpha(F)$ enables the model to consult the composite information in special pattern regions and get a more accurate anomaly map in Fig. 4 (e).

How multi-modal priors work? To investigate it, we visualize its impacts in Fig. 5. The multi-modal priors suppress responses to normal patterns in both anomaly-free and anomalous samples, *e.g.,* the chocolate on the "cookie" and the hollow on the "potato". This is mainly because the multi-modal priors contain normal information and are trained to help the student decoder restore anomaly-free features. Therefore, anomalous regions are highlighted and responses to normal patterns are mitigated after calculating pixel-wise feature similarity between the teacher and student networks.

Conclusion

We present a novel MMRD paradigm for anomaly detection, which integrates an auxiliary modality into RGB images for better detection. It uses a frozen multi-modal teacher encoder to generate multi-modal distillation targets for the learnable student decoder to restore. As a result, it achieves superior results on two multi-modal benchmarks.

Acknowledgments

This research is supported in part by National Natural Science Foundation of China (No. 61972157 and 72192821), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200), Shanghai Sailing Program (22YF1420300 and 23YF1410500), CCF-Tencent Open Research Fund (RAGR20220121) and Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

References

Bengs, M.; Behrendt, F.; Krüger, J.; Opfer, R.; and Schlaefer, A. 2021. Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain MRI. *International journal of Computer Assisted Radiology and Surgery*, 16(9): 1413–1423.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. Uninformed Students: Student-Teacher Anomaly Detection With Discriminative Latent Embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4182–4191. Computer Vision Foundation / IEEE.

Bergmann, P.; Jin, X.; Sattlegger, D.; and Steger, C. 2022. The MVTec 3D-AD Dataset for Unsupervised 3D Anomaly Detection and Localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 202–213. SCITEPRESS.

Bergmann, P.; and Sattlegger, D. 2023. Anomaly Detection in 3D Point Clouds using Deep Geometric Descriptors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2612–2622. IEEE.

Bonfiglioli, L.; Toschi, M.; Silvestri, D.; Fioraio, N.; and Gregorio, D. D. 2022. The Eyecandies Dataset for Unsupervised Multimodal Anomaly Detection and Localization. In *Proceedings of the Asian Conference on Computer Vision*, volume 13845, 459–475. Springer.

Chen, R.; Xie, G.; Liu, J.; Wang, J.; Luo, Z.; Wang, J.; and Zheng, F. 2023a. Easynet: An easy network for 3d industrial anomaly detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7038–7046.

Chen, X.; Han, Y.; and Zhang, J. 2023. A Zero-/Few-Shot Anomaly Classification and Segmentation Method for CVPR 2023 VAND Workshop Challenge Tracks 1&2: 1st Place on Zero-shot AD and 4th Place on Few-shot AD. *arXiv preprint arXiv:2305.17382*.

Chen, X.; Zhang, J.; Tian, G.; He, H.; Zhang, W.; Wang, Y.; Wang, C.; Wu, Y.; and Liu, Y. 2023b. CLIP-AD: A Language-Guided Staged Dual-Path Model for Zero-shot Anomaly Detection. *arXiv preprint arXiv:2311.00453*.

Deng, H.; and Li, X. 2022. Anomaly Detection via Reverse Distillation from One-Class Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9727–9736. IEEE.

Gou, J.; Sun, L.; Yu, B.; Du, L.; Ramamohanarao, K.; and Tao, D. 2022. Collaborative Knowledge Distillation via Multiknowledge Transfer. In *IEEE Transactions on Neural Networks and Learning Systems*. IEEE.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Gu, Z.; Liu, L.; Chen, X.; Yi, R.; Zhang, J.; Wang, Y.; Wang, C.; Shu, A.; Jiang, G.; and Ma, L. 2023. Remembering Normality: Memory-guided Knowledge Distillation for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16401–16409.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Horwitz, E.; and Hoshen, Y. 2022. An Empirical Investigation of 3D Anomaly Detection and Segmentation. *CoRR*, abs/2203.05550.

Hou, J.; Zhang, Y.; Zhong, Q.; Xie, D.; Pu, S.; and Zhou, H. 2021. Divide-and-Assemble: Learning Block-wise Memory for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8771–8780. IEEE.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141. Computer Vision Foundation / IEEE Computer Society.

Kim, T. S.; Jones, J. D.; and Hager, G. D. 2021. Motion Guided Attention Fusion to Recognize Interactions from Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13056–13066. IEEE.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9664–9674.

Liang, Y.; Zhang, J.; Zhao, S.; Wu, R.; Liu, Y.; and Pan, S. 2023. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing*.

Liu, H.; Liu, H.; Wang, Y.; Sun, F.; and Huang, W. 2022. Fine-grained multilevel fusion for anti-occlusion monocular 3d object detection. *IEEE Transactions on Image Processing*, 31: 4050–4061.

Liu, J.; Xie, G.; Chen, R.; Li, X.; Wang, J.; Liu, Y.; Wang, C.; and Zheng, F. 2023. Real3d-ad: A dataset of point cloud anomaly detection. *arXiv preprint arXiv:2309.13226*.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. V. 2022. Towards Total Recall in Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14298–14308. IEEE.

Rudolph, M.; Wandt, B.; and Rosenhahn, B. 2021. Same Same But DifferNet: Semi-Supervised Defect Detection with Normalizing Flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1906–1915. IEEE.

Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; and Wandt, B. 2023. Asymmetric Student-Teacher Networks for Industrial

Anomaly Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2591–2601. IEEE.

Rusu, R. B.; Blodow, N.; and Beetz, M. 2009. Fast Point Feature Histograms (FPFH) for 3D registration. In *2009 IEEE International Conference on Robotics and Automation*, 3212–3217. IEEE.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2021. Multiresolution Knowledge Distillation for Anomaly Detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14902–14912. Computer Vision Foundation / IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Wang, G.; Han, S.; Ding, E.; and Huang, D. 2021. Student-Teacher Feature Pyramid Matching for Anomaly Detection. In *British Machine Vision Conference*, 306. BMVA Press.

Wang, X.; Girshick, R. B.; Gupta, A.; and He, K. 2018. Non-Local Neural Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803. Computer Vision Foundation / IEEE Computer Society.

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020a. Deep Multimodal Fusion by Channel Exchanging. In *Advances in Neural Information Processing Systems*, 4835–4845. MIT.

Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; and Wang, C. 2023. Multimodal Industrial Anomaly Detection via Hybrid Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8032–8041.

Wang, Y.; Sun, F.; Lu, M.; and Yao, A. 2020b. Learning Deep Multimodal Feature Representation with Asymmetric Multi-layer Fusion. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3902–3910. ACM.

Xie, G.; Wang, J.; Liu, J.; Lyu, J.; Liu, Y.; Wang, C.; Zheng, F.; and Jin, Y. 2023a. Im-iad: Industrial image anomaly detection benchmark in manufacturing. *arXiv preprint arXiv:2301.13359*.

Xie, G.; Wang, J.; Liu, J.; Zheng, F.; and Jin, Y. 2023b. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. *arXiv preprint arXiv:2301.12082*.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339.

Zhang, J.; Chen, X.; Xue, Z.; Wang, Y.; Wang, C.; and Liu, Y. 2023a. Exploring Grounding Potential of VQA-oriented GPT-4V for Zero-shot Anomaly Detection. *arXiv preprint arXiv:2311.02612*.

Zhang, J.; Liu, R.; Shi, H.; Yang, K.; Reiß, S.; Peng, K.; Fu, H.; Wang, K.; and Stiefelhagen, R. 2023b. Delivering Arbitrary-Modal Semantic Segmentation. *CoRR*, abs/2303.01480.